The missing branches of the bee Tree of Life: addressing global Darwinian shortfalls and their drivers

Felipe Walter Pereira[1*], Matheus Lima Araujo[1], Anderson Lepeco[2], Bruno Ferreira Marques[1], Hugo Bampi[1], Lucas Jardim[3], Luísa Vareira[1], Luisa G. Carvalheiro[4], Thiago F. Rangel[4], José Alexandre F. Diniz Filho[4]

[1] Programa de Pós-Graduação em Ecologia e Evolução, Universidade Federal de Goiás, Goiânia, Brasil

[2] Laboratório de Entomologia, Departamento de Zoologia, Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil

[3] Laboratório de Macroecologia, Universidade Federal de Jataí, Goiás, Brasil

[4] Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Brasil

*Corresponding author: felip3walter@gmail.com

Orcid

FWP, https://orcid.org/0000-0002-0888-8804;

MLA, https://orcid.org/0000-0002-9111-725X;

AL, https://orcid.org/0000-0001-7467-5244;

BFM, https://orcid.org/0000-0002-6862-9564;

HB, https://orcid.org/0000-0001-7504-9894;

LJ, https://orcid.org/0000-0003-2602-5575;

LV, https://orcid.org/0000-0002-0917-9575;

LGC, https://orcid.org/0000-0001-7655-979X;

TFR, https://orcid.org/0000-0002-2001-7382;

JAFDF, https://orcid.org/0000-0002-0967-9684

Funding

FWP, MLA, AL, BFM, HB, and LV are granted by full PhD scholarships by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PROEX, grants 88887.910287/2023-00, 88887.903889/2023-00, 88887.208615/2025-00, 88887.991544/2024-0, 88887.660826/2022-00, and 88887.990433/2024-00, respectively). LJ is supported by the Tropical Water Research Alliance (TWRA)/FAPEG (conv. P&D&I TWRA/FAPEG 03/2023, proc. 202210267000536). LGC is funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grants 402743/2024-5 and 408201/2025-8). JAFDF and TFR have been continuously supported by CNPq productivity grants. This work on macroecology has been developed in the context of the National Institute of Science and Technology (INCT) in Ecology, Evolution, and Biodiversity Conservation, funded by CNPq (grants 465610/2014-5 / 409197/2024-6) and FAPEG (grant 201810267000023).

Acknowledgements

We thank Rodrigo B. Gonçalves and Tiago B. Quental for their insightful comments and discussions on a preliminary version of this research, presented at the I Congresso Brasileiro de Biologia Evolutiva. We also thank John Pickering and John S. Ascher for maintaining the Discover Life Apoidea Catalogue (bee species guide and world checklist), as well as Gabriel A. R. Melo for his invaluable work in maintaining the Moure's Bee Catalogue.

The missing branches of the bee Tree of Life: addressing global Darwinian shortfalls and their drivers

ABSTRACT

Understanding the Darwinian shortfall (i.e., the lack of knowledge about phylogenetic relationships) can help us to guide future biodiversity research and conservation efforts. Overcoming this shortfall is essential to develop robust strategies to preserve the Tree of Life while facing the ongoing biodiversity crisis. Here, we present the first global assessment of Darwinian shortfalls and their drivers in one of the main groups of pollinators, the bees. We built phylogenies for over 12000 bee species, combining the most comprehensive phylogeny and an algorithm with random solutions to insert missing lineages. The Darwinian shortfall was quantified as the Phylogenetic Diversity (PD) deficit, the ratio of inserted branch lengths, at the assemblage level. The highest shortfalls were identified in the Southern Hemisphere. Mean species range size and species richness were the strongest drivers, as smaller ranges and higher richness were associated with higher deficits. Per capita GDP was negatively associated with PD deficits, while population and road densities showed positive but weak effects. Sample completeness had a weaker effect, limited by missing occurrence data in many regions. Our findings underscore the need for integrative efforts combining taxonomy, data digitization, adequate research investments, and targeted sampling, especially in the Global South.

KEYWORDS: Anthophila, biodiversity, evolution, knowledge shortfalls, phylogenetic diversity

1. Background

A rare bimodal latitudinal gradient of taxonomic diversity is known and well described for bees, with the species richness peaking at dry, Mediterranean-type habitats outside the tropical zone. This was first theoretically discussed in light of the biogeography of bees (*1*) and more recently emphasized considering macroecological analysis (*2*). On the other hand, publicly available datasets of bees are biased towards North America and Europe, where knowledge about bee taxonomy and distribution is comparatively more consolidated, while well-known knowledge gaps are found for South America, Africa, and Asia (*2*, *3*). Additionally, richness-based accounting for diversity can often lead to biased biodiversity estimates, especially when considering the Linnean (i.e., discrepancy between described species and the number of all existing species (*4*, *5*)) and the Wallacean

shortfalls (i.e., lack of knowledge about geographic distribution of species (*4, 5*)). Biodiversity shortfalls have been demonstrated to hamper large-scale biodiversity assessments of bees, such as species decline and distribution patterns, even in Europe, where the bee fauna is relatively well-known due to a long tradition in melittology (*6*). Such shortfalls are expected to be even more pronounced in tropical regions, especially in Global South countries (*3, 7, 8*), as demonstrated for bees in Brazil (*9*).

Integrating evolutionary information is essential to better evaluate macroecological patterns, while identifying the impacts of biodiversity loss on the Tree of Life and, in some situations, partially overcoming Linnean shortfalls (*10*). However, knowledge about species taxonomy, geographic distribution, and evolutionary relationships remains incomplete, varying among taxa, and being unevenly distributed around the world, with more pronounced knowledge gaps for megadiverse taxa and regions (*11*, *12*, *13*, *14*). Thus, efforts to measure the Darwinian shortfall (i.e., the lack of evolutionary knowledge about phylogenetic relationships (*10*)) are crucial and might improve the rigor of evaluations of macroecological biodiversity patterns (*10, 15*), as demonstrated for European bees (*16*). Further, phylogenetic information can guide conservation priorities by identifying evolutionary distinct clades and regions that contribute disproportionately to better protect the Tree of Life (*14*, *16*). In addition, phylogenetic-based metrics are less sensitive to the Linnean shortfall and to the description of new species compared to those based exclusively on taxonomic richness (*10*); although the accuracy of diversification patterns descriptions may be positively affected by the addition of recent divergencies in phylogenies (*15*, *17*). Therefore, addressing and understanding the lack of knowledge about the evolutionary history of bees might lead to more effective strategies for further research and conservation (*6*, *10*).

Evolutionary relationships among bees have been better understood in the last decades, with huge efforts to clarify the origin, and diversification of major lineages (*18*, *19*, *20*, *21*, *22*). Recently, a phylogenomic and fossil-calibrated tree shed light on the origin and evolutionary history of bees, including 216 species representing all major lineages (*22*). Subsequently, a supermatrix phylogenetic tree was produced compiling all available phylogenetic data for bees, including 4,586 species, covering 22% of known species and 72% of genera – the most taxon-comprehensive phylogenetic tree currently available for bees (*23*). Presently, evolutionary relationships among bee families, subfamilies, and tribes are well known, remaining stable across different evaluations (*20*, *22*, *23*). However, the phylogenetic placement of nearly 80% of bee species remain unknown, evidencing unsolved uncertainties in relationships between and within most genera (*23*). This percentage indicates the extent of the large Darwinian shortfall observed for the group, although still unknown which clades and groups are predominantly affected by these shortfalls, where these lack of phylogenetic information are spatially concentrated, and what are their main drivers.

Phylogenetic lineage imputation (i.e., inserting missing species and lineages into a backbone phylogeny) is a feasible strategy to gather phylogenetic information from multiple sources (e.g., molecular phylogenies, taxonomy, and expert opinion), while accounting for the effect of uncertainty caused by incomplete phylogenetic knowledge. (*24*, *25*, *26*). Further, imputed phylogenies are useful to address the Darwinian shortfall in order to guide further research and conservation efforts (*14*, *26*). In this sense, the Darwinian shortfall can be quantified in terms of phylogenetic diversity (PD) deficit, the proportion of branch lengths that refers to imputed species in relation to the "complete" phylogeny, as proposed by Nakamura et al. (*26*). This approach provides a robust alternative to estimate our ignorance about the Tree of Life, as it relies on a measure of

133 missing branch lengths. It presents advantages by considering the proportion of missing
134 evolutionary history rather than just quantifying the proportion of species with publicly
135 available gene sequences – as often quantified in the literature (e.g. (*6*, *12*)).

136 Our main goal here is to address the global Darwinian shortfall of bees worldwide
137 by (i) comparing regional PD of bees worldwide before and after imputations (*26*); (ii)
138 highlighting clades in which there are more phylogenetically unrepresented species; (iii)
139 locating spatial gaps of phylogenetic information for bees; and (iv) identifying
140 macroecological and socioeconomic drivers of the lack of phylogenetic data for bees
141 worldwide. Thus, we expect that our results will provide a pathway to direct future efforts
142 to fill the gaps, increasing biodiversity knowledge and conservation of bees worldwide.

143

144 ## 2. Methods

145 ### *2.1. Occurrence data*

146 Global occurrence data was obtained following a recently published workflow
147 implemented in the *BeeBDC* R package (*3*). This workflow was proposed to aggregate,
148 standardize, add record-level flags for potential quality issues, and clean bee occurrence
149 data from multiple sources. Also, the authors provided a global bee occurrence dataset
150 combining more than 18 million uncleaned (6.9 million standardized and cleaned) bee
151 occurrences from multiple public repositories (e.g., GBIF, SCAN, iDigBio) and other
152 smaller data sources (i.e., non-public, private, or publicly inaccessible sources that shared
153 their data) – which are better detailed in the original publication.

154 Here we obtained the completely cleaned global dataset, publicly available and
155 last updated in February 2024 (*27*). For this dataset, the authors removed all records that
156 failed any of the filtering steps except for: (1) coordinate uncertainty based on a threshold
157 of ~1.1 km at the equator and (2) flagged old records collected before 1950. We have

158 decided to keep these records, as they provide valuable information on a macroecological
159 scale. This cleaned dataset comprises 6,785,860 occurrence records for 11,607 bee
160 species – meaning that occurrence data is openly available for only 55.4% of known
161 species (*28*).

162 We applied a spatial rarefaction of points for each species by identifying those
163 with the same first two decimal digits in their coordinates (~1.1 km at the equator) and
164 randomly keeping only one while discarding the others (resulting in 2,478,875 unique
165 records). This was a practical decision to remove very close points that introduced some
166 geometrical complications when defining geographical ranges in an initial trial (see next
167 section). Also, we removed records of *Apis mellifera*, as their present distribution mostly
168 results from human-driven actions (i.e., apiculture) and subsequent invasion events,
169 making it difficult to delineate its current native range. Finally, we removed exotic records
170 of species known to be (accidentally or intentionally) introduced outside their native
171 range, based on the most recent list available (*29*). In this latter process, six species for
172 which only exotic records are available were dropped. This dataset comprises 1,653,222
173 occurrence records for 11,600 bee species.

174 Additionally, we integrated a comprehensive database of bee occurrences in
175 Brazil (see (*9*) for further details), comprising over 500,000 records. This database
176 compiles digitized data from the public repositories GBIF and SIBBR
177 (https://www.sibbr.gov.br/), as well as the Moure's Bee Catalogue
178 (https://moure.cria.org.br/ (*30*)), which is the main reference for Neotropical bees. This
179 database also includes information from several entomological collections and from
180 digitized scientific articles. After removing duplicates (keeping only unique occurrences
181 that were absent in BeeBDC, and also removing close points with the same first two
182 decimal degrees, as above), we obtained 47,162 occurrences for the 1,965 bee species

known to occur in Brazil, of which 771 (39%) species (excluding synonyms based on the Discover Life Apoidea Catalogue (*28*)) were absent in the BeeBDC database.

A final step in data acquisition was to ensure that every species present in the backbone phylogeny ((*23*), see *2.4 Phylogenetic data* section for further detail) had occurrence data, as 482 of the 4586 species in the phylogeny were missing in the dataset with geographical records. For those species, we searched for occurrences in the primary literature by simply searching the species name in Google Scholar and obtaining occurrence records available from taxonomic studies. When no primary study about a species was found, we obtained occurrences available in the Discover Life Apoidea Catalogue (*28*). The entire process resulted in the addition of 3,133 records for all 482 species previously lacking distribution data (see supplementary material 1).

Our final dataset comprised over 1.7 million occurrences for 12,853 bee species – 61,5% of the 20,925 known valid species (*28*).

2.2. *Species geographical ranges*

We estimated the geographical ranges for each species, representing the extent of their occurrence records. For species with four or more occurrence records (n = 8,197), we estimated species ranges using alpha-hulls, as they reduce overprediction compared to convex hulls (i.e., minimum convex polygons) (*31*). Since different species require different alpha values (*32*), we fitted alpha-hulls for each species, starting with an alpha value of one and then increasing it incrementally by one until it returned a valid hull – encompassing at least 95% of occurrences (which allows the exclusion of dubious records too far from the others). The alpha-hulls algorithm is implemented in the *rangeBuilder* R package (*33*).

For some species, alpha-hulls could not be fitted (n = 9); for these species, as well as those with only three unique occurrences (n = 902), we used convex hulls plus a 100 km buffer instead. Finally, for species with one or two records (n = 3,754), we estimated species' ranges using 100 km buffers around each occurrence as a measure to address distribution uncertainty and data scarcity (*34*, *35*). Both convex hulls and buffers were created using the *sf* R package (*36*).

2.3. Presence-absence matrix

Species geographical ranges were gridded at a resolution of 100 x 100 km using the Behrmann equal-area projection. Species ranges were cropped to fit terrestrial landmasses, resulting in the exclusion of 187 ranges of species distributed in small islands or with small ranges near coasts. We then obtained a presence-absence matrix that displays all co-occurring species found in each grid cell for the 12,666 species. These procedures were carried out using the *EcoPhyloMapper* R package (*37*).

2.4. Phylogenetic data

We obtained "complete" phylogenetic trees for the 12,666 bee species by integrating the most taxon-comprehensive and up-to-date hypothesis available for the group ((*23*) available for download at BeeTree (<http://beetreeoflife.org/>)). This latter is based on a supermatrix approach, concatenating public genetic sequence data, including as the backbone the fossil-calibrated phylogenomic hypothesis of Almeida et al. (*22*). The resulting supermatrix phylogeny comprises 4,586 bee species, representing 23% of valid species and 82% of genera (*23*), and was used here as the backbone tree for the phylogenetic imputations of missing species.

We then obtained species-level phylogenetic trees using the framework proposed by Rangel et al. (24). The first step consisted in identifying a Phylogenetically Uncertain Taxon ("PUT", for a single taxon or clade, or "PUTs", for multiple taxa or clades), which are the species, groups of species, or even lower taxonomic groups of bees that are missing from the backbone (*23*). Subsequently, for each PUT, we defined their respective Most Derived Consensus Clade (MDCC) – corresponding to the node in the backbone tree that unequivocally contains each PUT (*24*).

To conservatively define the PUTs and MDCCs, we searched in the literature ("species name + phylogeny" in *Google Scholar*) for other phylogenetic studies that were not included in the original supermatrix tree (i.e., morphological phylogenies and recent molecular phylogenies published after the supermatrix tree). This search was replicated for each PUTs. This step provided valuable information to better define where each PUT would be imputed based on the most reliable information available (see supplementary material 2). For those PUTs lacking any hypothesis for phylogenetic placement, we defined the MDCCs as the clade corresponding to the highest taxonomic level available in the backbone tree (i.e., if other species from the same subgenus were available, we defined the subgenus as the MDCC; if no species from the same subgenus were available, then we defined the genus as the MDCC; and so on). Further, the resulting polytomies were solved by using an algorithm that applies random solutions for PUTs positions within their respective MDCCs ((*24*) but see (*38*) for detailed algorithm description). We simulated 1,000 trees accounting for uncertainty in imputations using an R package in development (Araújo et al., *in prep.*).

2.5. Measuring the Darwinian shortfall

First, species' geographical ranges were overlapped with 100 km Behrmann equal-area grid cells (herein assemblages) to obtain a presence-absence matrix accounting for species composition in each assemblage. We removed grid cells with less than three species to mitigate the impact of undersampled and unrealistic assemblages that might generate noise in the analysis. Second, the Phylogenetic Diversity (PD) was calculated as the sum of branch lengths (*39*) separately for each assemblage using both the backbone phylogeny (*23*) and the 1000 imputed phylogenies. Then, the Darwinian shortfall was measured as the Phylogenetic Diversity deficit (PD deficit), as proposed by Nakamura et al. (*26*):

$$PD_{deficit} = \frac{PD_{PUTs}}{PD_{total}}$$

Where PD_{PUTS} is the PD corresponding exclusively to inserted species in a given assemblage, while PD_{total} is the total PD from that assemblage. Finally, the mean values of PD deficits at the assemblage level were retained for further analysis of drivers of phylogenetic diversity, as well as the standard deviations of PD deficits to describe statistical uncertainty (supplementary material 3, figure s3). Therefore, the measured PD deficit represents the component of Darwinian shortfall led by the absence of phylogenetic information in the Tree of Life (*14*, *26*).

2.6. Drivers of the Darwinian shortfall

To identify drivers of the Darwinian shortfall at the assemblage level, we selected some general, widely used macroecological and socioeconomic variables. First, for biological potential predictors, we considered the following: (i) species richness, (ii) mean species range size, and (iii) corrected weighted endemism. The proxy of bee species richness is simply the species count for each assemblage based on the overlap of known species ranges. Mean species range sizes were calculated as the mean range size in km² of the bee

280 species occurring in the assemblage. Endemism was calculated as the sum of the inverse
281 of the range sizes of the species that occur in each cell divided by the total number of
282 species in each cell (*40*, *41*).

283 For socioeconomic variables we considered (i) population density for the year
284 2020 – gridded data at 30 arc seconds (~1 km) resolution (available from Center for
285 International Earth Science (*42*)); (ii) per capita gross domestic product (GDP) at 1 km²
286 resolution (*43*); and (iii) road density at 5 arc minutes (*44*). All socioeconomic variables
287 were aggregated to match the 100 km equal-area resolution. For population and road
288 density we extracted mean values, whereas for GDP we first summed gridded per capita
289 GDP within each 100 x 100 km grid cell and then divided this value by the total
290 population (population density * 10000 (area of each grid cell in km²)).

291 Additionally, we included sample completeness, as a measure of Wallacean
292 shortfall (*5*, *45*), as another potential driver of Darwinian shortfall. We quantified sample
293 completeness following the approach proposed by Chao et al. (*46*) using incidence data
294 for each assemblage. First, we created a presence-absence matrix for sub-grid cells of 10
295 km x 10 km resolution using the complete dataset of occurrence records (before the spatial
296 rarefaction by removing those with the same first two decimal coordinates digits). Then
297 incidences were quantified for each species present at each 100 km grid cell (i.e., the
298 frequencies of sub-grid cells occupied by each species), as incidence data are less
299 sensitive to aggregation and clustering found on abundance-based data (*46*, *47*). We
300 removed cells with fewer than 10 incidences as a filter rule to avoid unrealistic
301 extrapolations (*46*). Finally, we estimated sample completeness profiles for each 100 km
302 grid cells by estimating the slopes of incidence-based species accumulation curves (*46*).
303 We set q = 1 (i.e., the Hill number equivalent to the Shannon diversity index), as this
304 estimator accounts for the total number of incidences belonging to detected species,

without being too sensitive to infrequent species (as when q = 0, species richness) or favouring highly frequent species as in q = 2 (i.e., the Simpson diversity index). This approach properly quantifies sample completeness for incidence data when a species' weight is treated proportionally to its detection probability, as all individuals are weighted equally regardless of species identity (*46*). Sample completeness was computed using the *iNEXT* R package (*48*).

2.7. Modelling Phylogenetic Deficit

First, all variables, except for sample completeness (percentage), were log-transformed to improve normality. Then, the variables were standardized into Z-scores to allow comparability between effect sizes. Potential multicollinearity between variables was assessed by first fitting an ordinary least squares (OLS) regression model and then calculating variance inflation factor (VIF) values. As VIFs were moderate for all variables, ranging from 1.1 for endemism to 2.8 for GDP (supplementary material 3, table S1), we did not drop any variables. Residuals of the OLS were evaluated with Moran's I autocorrelation coefficient and a correlogram (supplementary material 3, figure S2). As significant spatial autocorrelation was found, we switched to simultaneous autoregressive (SAR) models (*49*), integrating spatial error into SAR models. We tested different neighbourhoods to define the list of weights, and we found that distance-based weights using inverse distance weighting (IDW) for neighbours in a radius of 3000 km were the most effective to reduce spatial autocorrelation. We fitted SAR error models for all combinations of predictors (*12*), considering only combinations of three or more variables – resulting in 99 candidate models.

We extracted model averaging based on Akaike information criterion (AIC) weights as model coefficients (i.e., slopes) across all candidate models (*50*). We selected

330 the minimum adequate model based on the AIC and, the, used Nagelkerke's pseudo-R^2
331 as a measure of explained variation (*51*). These models were fitted using the *spdep* R
332 package (*52*).

333 Finally, we performed an independent cross-species analysis to evaluate the effect
334 of range size on the probability of a species being phylogenetically known by fitting a
335 standard logistic regression of knowledge status (1 = presence, and 0 = absence in the
336 backbone phylogenetic tree) on square-root transformed range sizes (*12*). This model was
337 fitted using the *glm* function in base R.

338

339 **3. Results**

340 *3.1. Phylogenetic insertions*

341 Our phylogenies included 91% of all bee genera recognized (543 out of 598) after the
342 imputation of PUTs, with 72% of genera already present in the original backbone (figure
343 1a,c). The 543 genera comprise 12,666 bee species, over 60% of the 20,925 currently
344 described species (*28*). Out of all bee richness, 22% were already included in the
345 backbone phylogeny (*23*), and other 38% were imputed herein (figure 1b,d).

346 As expected, the phylogenetic imputations of PUTs increased the proportion of
347 species included per family more than the proportion of genera, since a high proportion
348 of genera – but a relatively low number of species – were already represented in the
349 backbone tree (figure 1). Halictidae and Megachilidae were the families for which
350 imputations most significantly increased the proportional representativeness of genera
351 (figure 1a), while the distribution of species proportions was more evenly spread across
352 families (figure 1b).

353

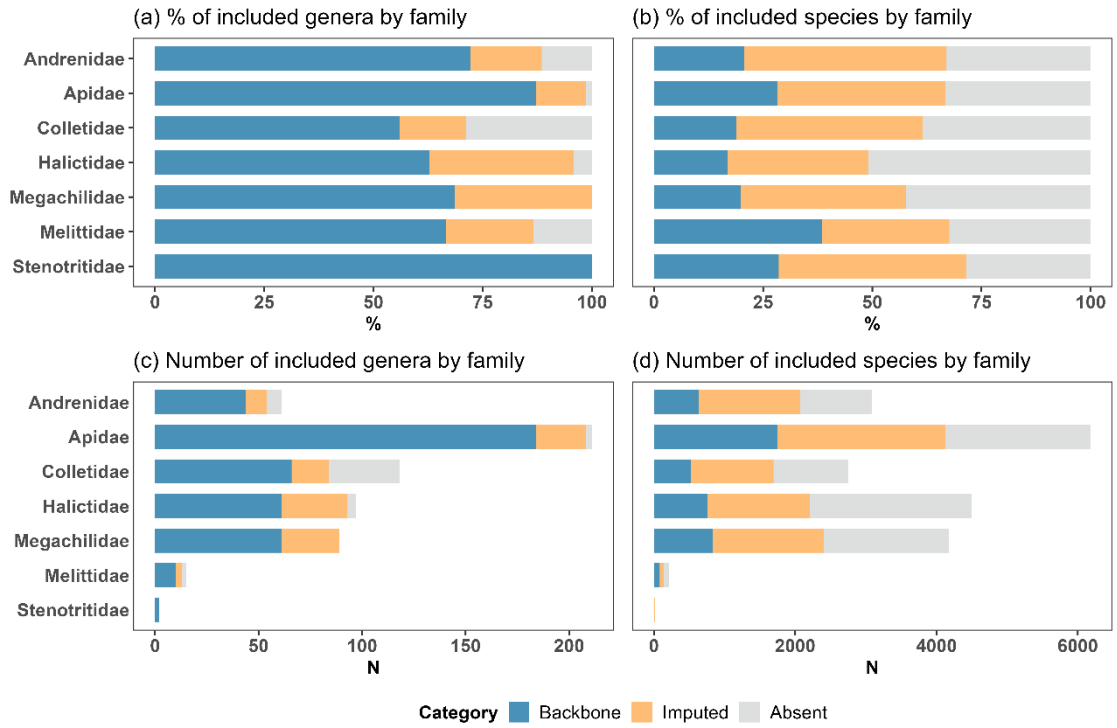


Figure 1. Bar plots summarize the percentage of (a) genera and (b) species, as well as the absolute number of (c) genera and (d) species, included in the phylogeny for each bee family. Blue represents genera or species already present in the backbone phylogeny (*23*), orange represents the portion imputed into the phylogeny, and grey represents the portion of genera or species absent from the final phylogeny due to lack of geographical information.

Coverage of genera and species included in the phylogeny after imputations varied among the seven bee families, ranging from 71% up to 100% for genera (figure 1a) and from 49% to 71% for species (figure 1b). The bee families with higher representativeness of genera were Megachilidae (n = 89) and Stenotritidae (n = 2), both with all genera included in the phylogeny after imputations of PUTs. The most diverse bee family, Apidae, was represented by 99% of the valid genera. On the other hand, Colletidae was the family with the lowest genera representativity, with 71%. As expected, Stenotritidae was the family with the highest proportion of species included in the phylogeny (71%), as this is the least diverse family with only 21 valid species, followed by Melittidae (the second least diverse family) with 68% of species. Finally, the bee family with the lowest proportion of species included in the phylogeny was Halictidae (49%), the second most diverse bee family (figure 1d).

3.2. Sample completeness and the Wallacean shortfall

Higher sample completeness values were found in the midwestern and western regions of the USA, indicating lower Wallacean shortfalls. Northern portions of Mexico also showed high sample completeness. Interestingly, moderate to high completeness was quantified for assemblages along the southeastern and eastern coasts of Brazil, what may be due to the inclusion of additional occurrence records for Brazil. However, most parts of Brazil – especially the central and northern regions – still lack information on bee distributions. The same is observed across much of South America, where few assemblages have available occurrence records (figure 2). Overall, occurrence data deficiency remains predominant across most regions, except for North America and Western Europe (figure 2).

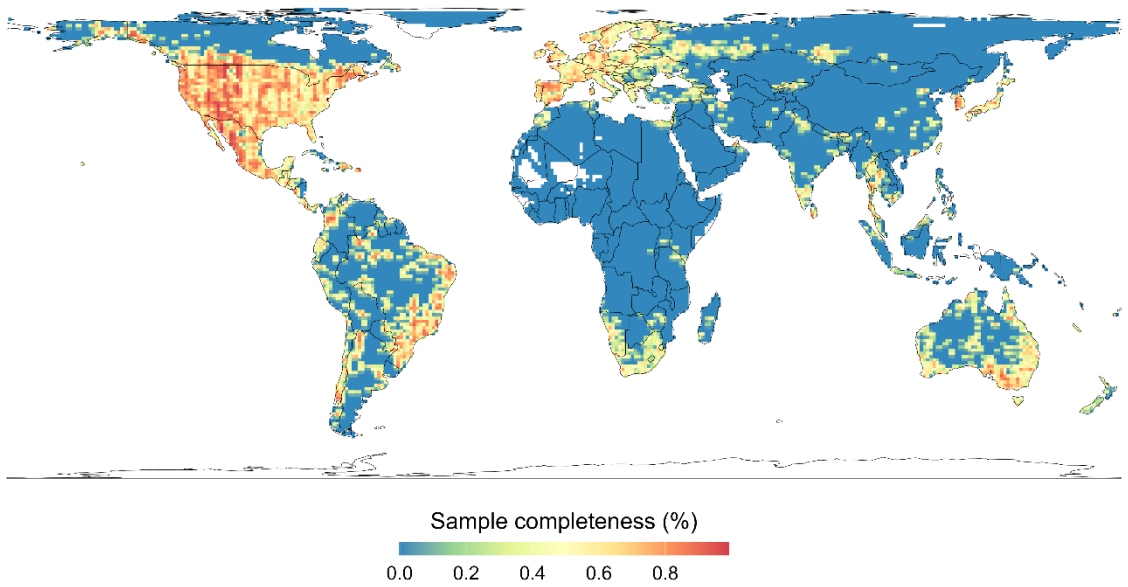


Figure 2. Incidence-based sample completeness of wild bees estimated for 100 x 100 km equal-area grid cells. Sample completeness was estimated using q = 1 (equivalent to Shannon diversity); see *Methods* for further details. Lower sample completeness indicates higher Wallacean shortfall.

3.3. The Darwinian shortfall in phylogenetic knowledge of wild bees

The mean values of PD deficits at the assemblage-level (from the 1,000 replicates) were a consistent measure of the Darwinian shortfall, as standard deviations were extremely low – with a maximum SD of 0.013 (supplementary material, figure S3). Substantial differences between PD values measured using only the backbone tree (figure 3a) and after the imputation of PUTs (figure 3c) were particularly evident in the USA and Mexico, southeastern South America, southern Africa, and eastern and western coasts of Australia. This pattern is even more pronounced when considering only the branch lengths of PUTs inserted into the backbone tree (figure 3b), where longer branch lengths were added, indicating that major lineages from these regions lack phylogenetic information. In contrast, PUTs from western Europe contributed relatively little to the PD of assemblages, suggesting that most lineages (i.e., most tribes and genera) from these regions are already represented in existing molecular phylogenies.

Higher PD deficits were observed across the Neotropics, Afrotropic, western and eastern coast of Australia, New Guinea, and southwestern USA (figure 3d). Some of these regions were expected to exhibit higher PD deficits due to a combination of high bee diversity and limited species representation in the backbone phylogeny (as for southeastern South America and southern Africa). Conversely, lower PD deficits were found in most regions of Europe. Despite being one of the countries with good representation of bees in the backbone phylogeny, moderate to high PD deficits were found for the USA, suggesting that substantial phylogenetic knowledge remains to be uncovered even in regions known for their high bee richness (e.g., the southwestern USA). Additionally, lower PD deficits were observed in regions where bee diversity is naturally lower, such as the high latitudes of the Northern Hemisphere (figure 3d).

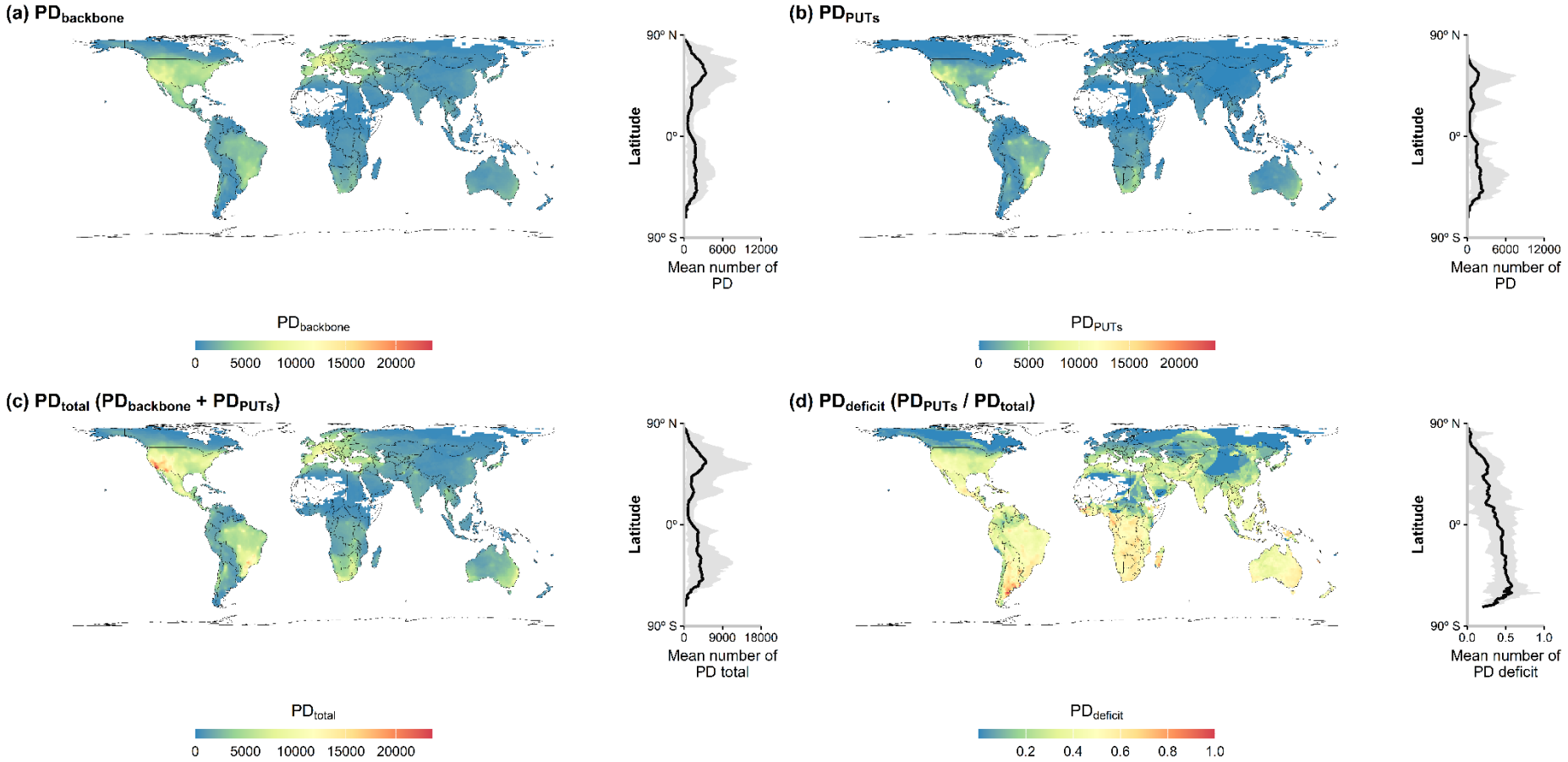


Figure 3. Phylogenetic diversity (PD) and Darwinian shortfall (PD deficit) of wild bees worldwide. (a) Phylogenetic diversity measured using only the backbone tree; (b) Phylogenetic diversity corresponding to the branch lengths of PUTs inserted into the backbone; (c) PD measured using the final phylogenetic tree after the imputation of PUTs; (d) PD deficit, representing the Darwinian shortfall. Latitudinal distribution curves are shown on the right side of each map. For (b–d), we are using means over the 1,000 imputed phylogenies. Pixels in white represent cells without any known species ranges overlapping.

3.4. Drivers of the Darwinian shortfall

The best-fitting model to explain Darwinian shortfalls of wild bees worldwide included all variables except endemism as predictors, explaining 82% of the variance in the PD deficit (pseudo-R^2 = 0.825). The other model with $\Delta AIC < 2$ included all the seven variables and with basically the same pseudo-R^2 (equal to 0.825). Across all models, mean range size had the strongest negative effect. Sample completeness and GDP also had negative effects, although with shallower standardized slopes. Endemism has a very weak negative effect, with almost flat slope. On the other hand, species richness had the strongest positive effect. Population density and road density also had positive effects, but with shallower slopes (table 1).

Table 1. Standardized slopes (z) of predictors of bee PD deficits included in all candidate SAR error models. Model averaged z values, as well as 95% interval standard errors (SE), were obtained from AIC-weighted averaging across all candidate models and then standardized with PD deficit and the predictors. Best model z values refer to the best-fitting, minimum adequate model. Best model's pseudo-R^2 and AIC weight are also presented. Detailed information is presented in supplementary material 3, tables S2 and S3.

Predictor	**Model averaged z**	**SE**	**Best model z**
Species richness	0.2259	0.0020	0.2276
Sample completeness	-0.0469	0.0003	-0.0472
Mean range size	-0.3513	0.0045	-0.3475
Population density	0.0985	0.0023	0.0970
per capita GDP	-0.0765	0.0020	-0.0749
Road density	0.0497	0.0003	0.0494
Endemism	-0.0107	0.0000	NA
Peudo-R2	-	-	0.825
AIC weight	-	-	0.58

Species richness and sample completeness were the predictors with the highest importances across models (>0.99), followed by mean range size (0.58) and population density (0.41). On the other hand, GDP, road density, and endemism (<0.01) were identified

with lower importances across models (figure 4). Candidate models with a similar formula to the best model, but either removing species richness or mean species range had slightly smaller R^2 compared to the full model (pseudo-R^2 = 0.793 for the one excluding mean range, and pseudo-R^2 = 0.811 for the one excluding species richness; see supplementary material 3, table S3).

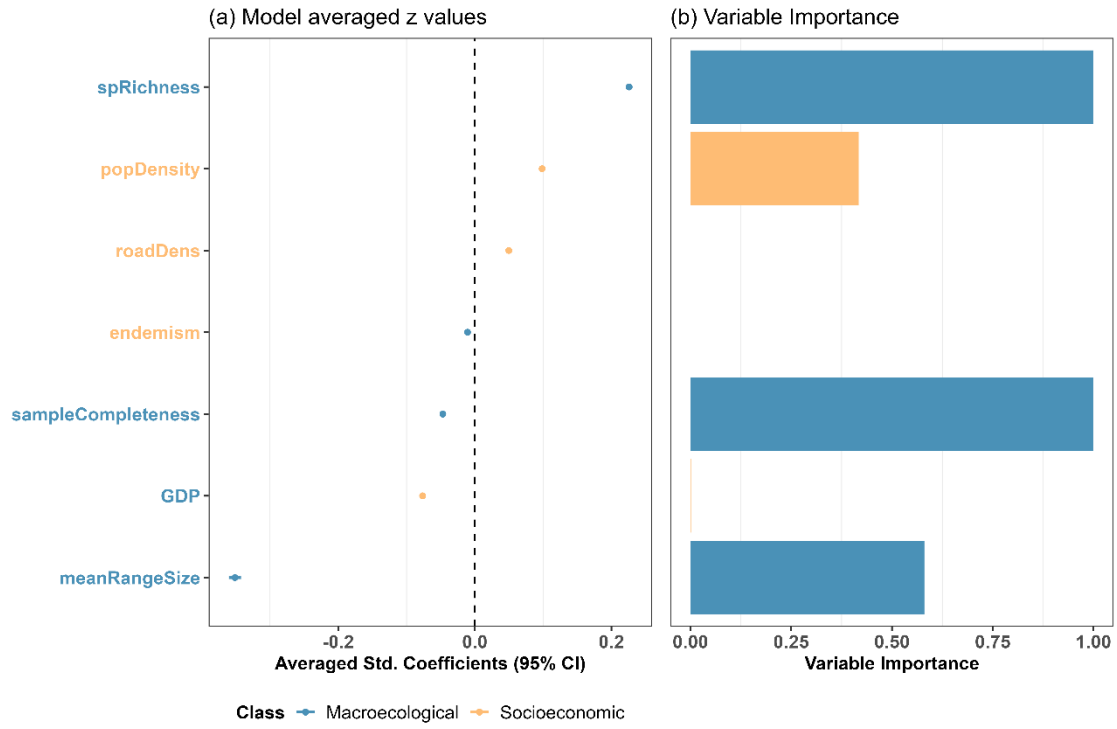
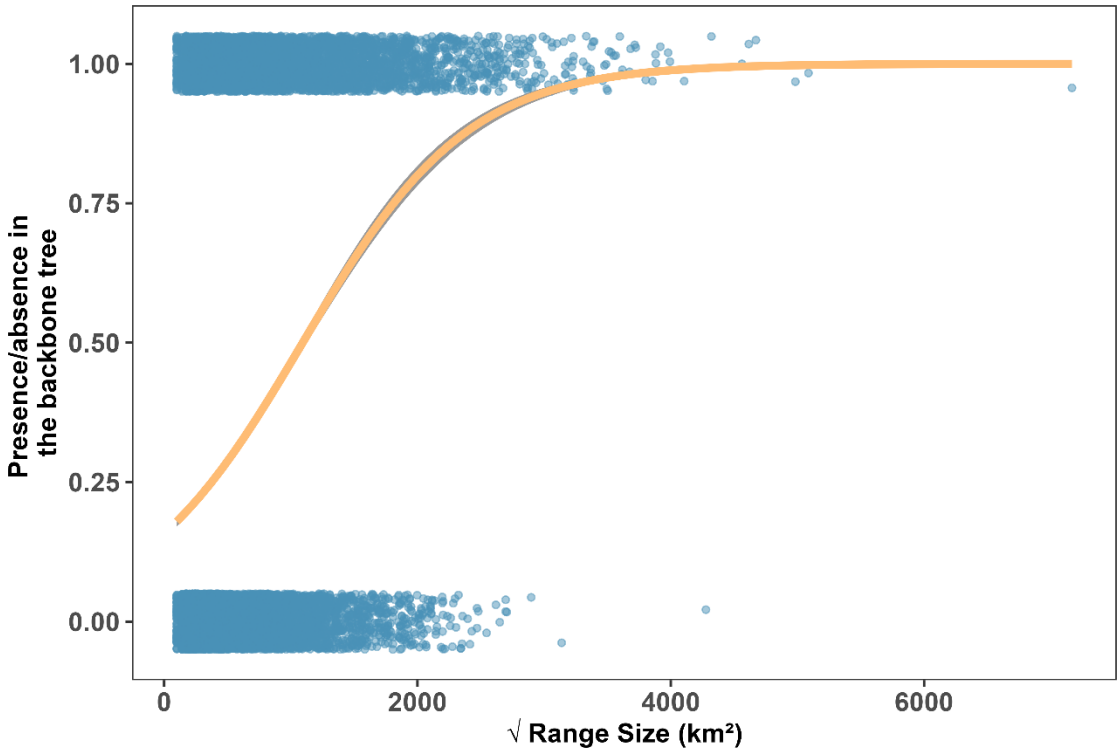


Figure 4. (a) Averaged z values with 95% confidence intervals, and (b) variable importance from model averaging across all candidate models for the included predictors. Averaged standardized coefficients, as well as 95% Confidence Intervals (CI), were obtained from AIC-weighted averaging across all candidate models and then standardized with PD deficit and the predictors. Variable importance was calculated as the sum of weights of models containing the variable. Blue represents macroecological variables, while orange represents socioeconomic variables.

Although our model explained over 80% of the variance in the PD deficit, it could not completely remove the spatial autocorrelation (supplementary material 3, figures S4–S6). This is especially due to some regions exhibiting lower species richness (supplementary material 3, figure S7), lower PD and relatively high PD deficits values, and the Andes, where

458 a relatively high richness and lower PD deficits are found (figure 3c,d). On the other hand,
459 the model was effective in reducing spatial autocorrelation in all the other regions where bee
460 data is more abundant and more consistent values of PD are found (figure 3c).

461 Finally, species with larger range sizes had higher probabilities of being
462 phylogenetically sampled (i.e., included in the backbone tree), with an increase of 0.15% per
463 unit increase in square-root range size (estimate = 0.00153 ± 0.000041, z = 37.34, $p < 0.001$),
464 as estimated with logistic regression (figure 5).



465

466 **Figure 5.** Logistic regression of phylogenetic knowledge status of a species (1 = presence, and 0 =
467 absence in the backbone phylogenetic tree) and their square-root transformed range size. Probability
468 of being phylogenetically known is indicated by the orange curve.
469

470

471 **4. Discussion**

Here, we present the first global-scale assessment of the Darwinian shortfall in wild bees, based on publicly available occurrence data and phylogenetic imputations into a broadly comprehensive phylogeny. Although this approach incorporated ~60% of known bee species, over 90% of genera were successfully included. As previously demonstrated, bee data availability is biased toward North America and Western Europe, especially in terms of sample completeness (*2*, *3*) and phylogenetically addressed species (*22*, *23*). Phylogenetic imputations allowed us to demonstrate a substantial increase in PD in southern continents. Consequently, higher Darwinian shortfalls, in terms of PD deficits, were found in these regions, highlighting that they harbour substantial evolutionary diversity of bees that has yet to be documented. We found that Darwinian shortfalls in wild bees, in general, increase in assemblages with higher estimates of species richness due to larger numbers of missing species, although this result is far from homogenous across the globe. Additionally, our results show that PD deficits decrease with higher mean species range size and sample completeness. Two of the socioeconomic factors, population and road densities, are associated with higher PD deficits, though with weaker effects. Finally, species with larger range sizes are more likely to be included in a phylogeny than those with smaller range sizes.

Furthermore, we demonstrate that the bimodal latitudinal taxonomic diversity pattern of bees (*1*, *2*) is followed by a similarly shaped phylogenetic diversity gradient (figure 2). This pattern can be clearly visualized from the backbone tree, being reproduced in the analysis based on the imputed phylogeny. In addition, it is worth mentioning that we found the peak of phylogenetic diversity in the Northern Hemisphere to be only slightly higher than that in the Southern Hemisphere. This is a much smaller difference than that shown for bee species richness (*2*). Even though our imputations successfully incorporated many missing branches from Southern Hemisphere lineages, a comparatively larger deficit of phylogenetic

lineage sampling in the Southern Hemisphere – as evidenced by the peaks of PD deficits. In this sense, we can expect equal or even higher phylogenetic diversity in the Southern Hemisphere than in the North as we overcome the Darwinian shortfall. From a historical biogeography perspective, this is not unexpected, given that many early-diverging lineages representing long branches can be found in South America and Africa, as those regions housed the earliest steps of bee evolution (*1*, *21*, *22*).

4.1. Taxonomic coverage

Taxonomic representation of bee species in the backbone phylogeny is uneven across families at both the genus and species levels (*23*). Although our phylogenetic imputations improved overall coverage, some families remained comparatively more well represented. Regarding genera, Colletidae were proportionally the least represented, leaving fine-scale relationships within its clades unresolved (see (*53*, *54*)). At the species level, Halictidae, the second most diverse family, remained poorly represented, with fewer than 50% of the known species included in the imputed phylogeny. This is particularly evident in the species-rich and widely distributed genus *Lasioglossum*, which comprises more than 1,800 described species (*28*), yet still presents major uncertainties regarding relationships within and among subgenera (*23*, *55*, *56*). Similar issues are found in *Andrena* (Andrenidae), although substantial progress has been made in the past decade (e.g. (*57*, *58*)). While a group-by-group evaluation is beyond the scope of this study, these examples illustrate persistent gaps in phylogenetic knowledge of bees. Future research expanding taxonomic representation in these key groups is expected to refine their phylogenetic relationships and clarify their evolutionary histories.

4.2. Data availability

The pervasive impacts of the Wallacean shortfall on bees are stronger in Global South countries, as recently demonstrated for Brazil (*9*), where almost 60% of the country's land area is devoid of distribution records. Although some of these regions truly represent understudied areas where little or nothing is known about their bee faunas (*59*), important distribution data may exist for many of them but remain inaccessible or undigitized (*9*). This issue is not exclusive to Brazil but rather a major problem across most regions of the world (*3*, *7*, *8*). Data inaccessibility also affects inferences even for the relatively more well-known bee faunas of Western Europe and the contiguous USA(*6*, *60*).

Despite increasing efforts in data digitization of bees in Western Europe, moderate to high Wallacean shortfalls are still evident throughout the region, as also noted in a previous analysis (*6*). Lower completeness values were also observed in most of Africa, where Wallacean shortfalls are even more pronounced, given the widespread scarcity of bee distribution data across the continent. Similarly, bee occurrence data is sparse throughout Asia, except for Japan and South Korea. Australia presents moderate sample completeness for assemblages near the coasts – particularly in the east – while central regions are mostly devoid of data, likely due to the dominance of desert areas.

4.3. The Darwinian shortfall in wild bees

Higher degrees of Darwinian shortfall underestimation are expected for regions where bee research has been historically less developed. This is of especial relevance given that occurrence data is not publicly available for nearly 40% of bee species. In addition, the range for part of the sampled species is presumably underestimated, since they may spread to

regions without extensive sampling efforts. The lack of digitization of data that has been already sampled may also hinder the evaluation of the Darwinian shortfall in these regions, especially in the tropics (*2*, *3*, *23*). Furthermore, expressive Linnean shortfall is also evident in these areas, where a significant proportion of species remain undescribed and major additions are expected in the future (*8*). Nevertheless, the present identification of major Darwinian shortfalls and their drivers is relatively sound, as it relies on information available for over 90% of known bee genera worldwide. Moreover, these findings align with previous studies demonstrating that the tropics are overall the least represented in molecular databases, paramount for building robust phylogenetic hypotheses (*12*, *61*). The relatively lower representation of tropical species is expected to have a major impact on estimates even for taxa that are more diverse in mid latitudes, as is the case of bees.

4.4. Drivers of the Darwinian shortfall

Species with more widespread distributions are more likely to be detected and subsequently addressed in phylogenetic investigations (*12*, *62*). Species richness was the next most influential factor, with PD deficits increasing in speciose areas. This result was expected, as larger Darwinian shortfalls might be expected in species-rich regions due to the given relationship between PD metrics and richness (*63*). Furthermore, it is important to note that regions with higher estimates of species richness may also be the ones with lower Linnean shortfalls, while other regions presenting lower richness may be a reflect of incomplete knowledge rather than a biological process (*8*).

564 Per capita GDP negatively affects PD deficit (though with smaller effect than mean
565 range size and richness), indicating that regions with higher incomes also have better-
566 understood clades, likely related to larger research expenditures (*12, 64*). The gap between
567 the Global North and South is even more pronounced when considering molecular
568 phylogenies. Even though access to molecular data has increased in recent decades due to the
569 overall reduction in the cost of DNA sequencing (*65*), it is still unavailable for many research
570 groups in megadiverse regions (*66, 67*). Nonetheless, GDP values alone may not translate
571 the effort in studying a particular region. Biodiversity research efforts in Global South
572 regions are frequently done by researchers from the Global North, thus reflecting a
573 geopolitical process (*12, 68, 69, 70*).

574 Although the slope is shallow, the PD deficit also decreases with higher sample
575 completeness, suggesting that well-sampled assemblages are more likely to have more
576 represented lineages in terms of phylogenetic knowledge. However, sample completeness
577 could not be estimated for many assemblages, especially in Asia, Africa, and South America
578 (figure 2). This limitation may explain the small effect of this predictor, as those cells were
579 treated as having zero completeness.

580 In contrast, PD deficit increases with population density and road density. The
581 positive effect of population density is expected in regions where high human populations
582 coincide with lesser-known bee faunas (e.g., southern and southeastern Asia). The
583 relationship with road density is less straightforward, since accessibility is expected to reduce
584 deficits (*71*). However, it is possible that regions that are inaccessible have substantial
585 Linnean shortfall (*8*), which bias the Darwinian shortfall to lower values. Still, road density
586 showed only marginal effects and may influence Darwinian shortfalls more strongly at

broader spatial scales (*12*, *14*). Finally, the negative but almost flat association between PD deficit and endemism contradicts the expectation that species with smaller ranges are less likely to have been studied, as reflected by mean range size. Even though this relation is weak, it can potentially reflect focused efforts of researchers to sample areas known to house highly unique bee faunas – something essential for lineage representativity in phylogenetic studies.

4.5. Conclusion and future perspectives

This study provides the first comprehensive evaluation of Darwinian shortfalls worldwide, highlighting both progress and our limitations in understanding the bee Tree of Life. While our results are robust, encompassing over 90% of bee genera, persistent biases in occurrence records and limited data digitization indicate that knowledge gaps remain particularly severe and underestimated in tropical and Global South regions. These areas often coincide with highly threatened biodiversity hotspots, underscoring the urgent need for increased sampling and conservation efforts to better understand and protect them (*72*). Addressing Darwinian shortfalls in wild bees, as well as other biodiversity knowledge gaps, will require effective broad-scale data sharing from collections and museums (*3*, *9*, *64*, *73*), alongside sustained investment in fieldwork and taxonomic expertise (*74*). Finally, strengthening international collaboration will be critical to ensure that the evolutionary history of bees is adequately documented and can inform effective conservation strategies.

Data accessibility

609 The data and code used for the analysis of this manuscript is available in a Figshare
610 repository, which can be accessed with a private anonymous link created for the reviewing
611 process (https://figshare.com/s/694071403bcd34143484).

612 **REFERENCES**

613 1. Michener CD. 1979 Biogeography of the bees. *Ann. Missouri Bot. Gard.* 66, 277–
614 347. (https://doi.org/10.2307/2398833)

615 2. Orr MC, Hughes AC, Chesters D, Pickering J, Zhu CD, Ascher, JS. 2021 Global
616 Patterns and Drivers of Bee Distribution. *Curr. Biol.* 31, 451–458.
617 (https://doi.org/10.1016/j.cub.2020.10.053)

618 3. Dorey JB *et al.* 2023 A globally synthesised and flagged bee occurrence dataset and
619 cleaning workflow. *Sci. Data* 10, 747. (https://doi.org/10.1038/s41597-023-02626-w)

620 4. Lomolino MV. 2004 Conservation biogeography. In: Lomolino MV, Heaney LR.
621 (eds) *Frontiers of Biogeography: New Directions in the Geography of Nature*, 293–
622 96. Sunderland, MA: Sinauer

623 5. Hortal J, de Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ. 2015
624 Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annu. Rev. Ecol.*
625 *Evol. Syst.* 46, 523–549. (https://doi.org/10.1146/annurev-ecolsys-112414-054400)

626 6. Marshall L, Leclercq N, Carvalheiro LG, Dathe HH, Jacobi B, Kuhlmann M, Potts
627 SG, Rasmont P, Roberts SP, Vereecken NJ. 2024 Understanding and addressing
628 shortfalls in European wild bee data. *Biol. Conserv.* 290, 110455.
629 (https://doi.org/10.1016/j.biocon.2024.110455)

630 7. Archer CR, Pirk CWW, Carvalheiro LG, Nicolson SW. 2014 Economic and
631 ecological implications of geographic bias in pollinator ecology in the light of
632 pollinator declines. *Oikos* 123, 401–407. (https://doi.org/10.1111/j.1600-
633 0706.2013.00949.x)

634 8. Dorey J, Gilpin A, Johnston N, Esquerré D, Hughes A, Ascher J, Orr M. 2025 How
635 many bee species are there? A quantitative global estimate. *Research Square*
636 (preprint). (https://doi.org/10.21203/rs.3.rs-6372769/v1)

637 9. Carvalheiro LG, Cordeiro GD, Marques BF, Menezes PP, Consorte PM, Gianinni
638 TC. 2025 Challenges for quantifying knowledge shortfalls on tropical pollinators in
639 the face of global environmental change – Brazilian bees as a case study. *Sociobiology*
640 72, e11276. (https://doi.org/10.13102/sociobiology.v72i2.11276)

641 10. Diniz-Filho JAF, Loyola RD, Raia P, Mooers AO, Bini LM. 2013 Darwinian
642 shortfalls in biodiversity conservation. *Trends Ecol. Evol.* 28, 689–695.
643 (https://doi.org/10.1016/j.tree.2013.09.003)

644 11. Carvalho FG, Duarte L, Seger GDS, Nakamura G, Guillermo-Ferreira R, Cordero-
645 Rivera A, Juen L. 2022 Detecting Darwinian Shortfalls in the Amazonian Odonata.
646 *Neotrop. Entomol.* 51, 404–412. (https://doi.org/10.1007/s13744-022-00961-y)

12. Rudbeck AV, Sun M, Tietje M, Gallagher RV, Govaerts R, Smith SA, Svenning J, Eiserhardt WL. 2022 The Darwinian shortfall in plants: phylogenetic knowledge is driven by range size. *Ecography 2022*, e06142. (https://doi.org/10.1111/ecog.06142)

13. Šmíd J. 2022 Geographic and taxonomic biases in the vertebrate tree of life. *J. Biogeogr.* 49, 2120–2129. (https://doi.org/10.1111/jbi.14491)

14. Soares BE, Nakamura G, Freitas TM, Richter A, Cadotte, M. 2023 Quantifying and overcoming Darwinian shortfalls to conserve the fish tree of life. *Biol. Conserv.* 285, 110223. (https://doi.org/10.1016/j.biocon.2023.110223)

15. Diniz-Filho JAF, Jardim L, Guedes JJ, Meyer L, Stropp J, Frateles LEF, Pinto RB, Lohmann LG, Tessarolo G, Carvalho CJ, Ladle RJ, Hortal J. 2023 Macroecological links between the Linnean, Wallacean, and Darwinian shortfalls. *Front. Biogeogr.* 15, e59566. (https://doi.org/10.21425/f5fbg59566)

16. Leclercq N, Marshall L, Caruso G, Schiel K, Weekers T, Carvalheiro LG, Dathe HH, Kuhlmann M, Michez D, Potts SG, Rasmont P, Roberts SPM, Smagghe G, Vandamme P, Vereecken NJ. 2023 European bee diversity: Taxonomic and phylogenetic patterns. *J. Biogeogr.* 50, 1244–1256. (https://doi.org/10.1111/jbi.14614)

17. Frateles LEF, Tavares GRG, Nakamura G, Silva NJ, Terribile LC, Diniz-Filho JAF. 2025 The interaction between the Linnean and Darwinian shortfalls affects our understanding of the evolutionary dynamics driving diversity patterns of new world coralsnakes. *J. Biogeogr.* 52, 42–54. (https://doi.org/10.1111/jbi.15014)

18. Danforth BN, Sipes S, Fang J, Brady SG. 2006 The history of early bee diversification based on five genes plus morphology. *Proc. Natl. Acad. Sci. USA* 103, 15118–15123. (https://doi.org/10.1073/pnas.0604033103)

19. Cardinal S, Danforth BN. 2013 Bees diversified in the age of eudicots. *Proc. R. Soc. B.* 280, 20122686. (https://doi.org/10.1098/rspb.2012.2686)

20. Danforth BN, Cardinal S, Praz C, Almeida EAB, Michez D. 2013 The impact of molecular data on our understanding of bee phylogeny and evolution. *Annu. Rev. Entomol*. 58, 57-78. (https://doi.org/10.1146/annurev-ento-120811-153633)

21. Hedtke SM, Patiny S, Danforth BN. 2013 The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. *BMC Evol. Biol.* 13, 138. (https://doi.org/10.1186/1471-2148-13-138)

22. Almeida EA, Bossert S, Danforth BN, Porto DS, Freitas FV, Davis CC, Murray EA, Blaimer BB, Spasojevic T, Ströher PR, Orr MC, Packer L, Brady SG, Kuhlmann M, Branstetter MG, Pie MR. 2023 The evolutionary history of bees in time and space. *Curr. Biol.* 33, 3409–3422. (https://doi.org/10.1016/j.cub.2023.07.005)

683 23. Henríquez-Piskulich P, Hugall AF, Stuart-Fox D. 2024 A supermatrix phylogeny of
684 the world's bees (Hymenoptera: Anthophila). *Mol. Phylogenet. Evol.* 190, 107963.
685 (https://doi.org/10.1016/j.ympev.2023.107963)

686 24. Rangel TF, Colwell RK, Graves GR, Fučíková K, Rahbek C, Diniz-Filho JAF. 2015
687 Phylogenetic uncertainty revisited: Implications for ecological analyses. *Evolution*
688 69, 1301–1312. (https://doi.org/10.1111/evo.12644)

689 25. Chang J, Rabosky DL, Alfaro ME. 2020 Estimating Diversification Rates on
690 Incompletely Sampled Phylogenies: Theoretical Concerns and Practical Solutions.
691 *Syst. Biol.* 69, 602–611. (https://doi.org/10.1093/sysbio/syz081)

692 26. Nakamura G, Richter A, Soares BE. 2021 FishPhyloMaker: An R package to generate
693 phylogenies for ray-finned fishes. *Ecol. Inform.* 66, 101481.
694 (https://doi.org/10.1016/j.ecoinf.2021.101481)

695 27. Dorey JB *et al.* 2023 Dataset for a globally synthesised and flagged bee occurrence
696 dataset and cleaning workflow. Flinders University. Dataset.
697 (https://doi.org/10.25451/flinders.21709757)

698 28. Ascher JS, Pickering J. 2024 Discover Life bee species guide and world checklist
699 (Hymenoptera: Apoidea: Anthophila).
700 (http://www.discoverlife.org/mp/20q?guide=Apoidea_species)

701 29. Russo L. 2016 Positive and Negative Impacts of Non-Native Bee Species around the
702 World. *Insects 7*, 69. (https://doi.org/10.3390/insects7040069)

703 30. Moure JS, Urban D, Melo GAR. 2024 Catalogue of Bees (Hymenoptera, Apoidea) in
704 the Neotropical Region. (https://moure.cria.org.br/catalogue)

705 31. Burgman MA, Fox JC. 2003 Bias in species range estimates from minimum convex
706 polygons: implications for conservation and options for improved planning. *Anim.*
707 *Conserv.* 6, 19–28. (https://doi.org/10.1017/s1367943003003044)

708 32. Meyer L, Diniz-Filho JAF, Lohmann LG. 2018 A comparison of hull methods for
709 estimating species ranges and richness maps. *Plant Ecol. Divers.* 10, 389–401.
710 (https://doi.org/10.1080/17550874.2018.1425505)

711 33. Rabosky DA, Cox C, Rabosky D, Title P, Holmes I, Feldman A, McGuire J. 2016
712 Coral snakes predict the evolution of mimicry across New World snakes. *Nat.*
713 *Commun.* 7, 11484. (https://doi.org/10.1038/ncomms11484)

714 34. Pokorny T, Loose D, Dyker G, Quezada-Euán JJG, Eltz T. 2014 Dispersal ability of
715 male orchid bees and direct evidence for long-range flights. *Apidologie* 46, 224–237.
716 (https://doi.org/10.1007/s13592-014-0317-y)

35. Vélez D, Vivallo F. 2024 Key areas for conserving and sustainably using oil-collecting bees (Apidae: Centridini, Tapinotaspidini, Tetrapediini) in the Americas. *J. Insect Conserv.* 28, 1247–1263. (https://doi.org/10.1007/s10841-024-00620-0)

36. Pebesma E, Bivand R. 2023 *Spatial Data Science: With Applications in R*. Chapman and Hall/CRC. (https://doi.org/10.1201/9780429459016)

37. Title P, Swiderski D, Zelditch M. 2022 EcoPhyloMapper: an R package for integrating geographic ranges, phylogeny, and morphology. *Methods Ecol. Evol.* 13, 1912-1922. (https://doi.org/10.1111/2041-210X.13914)

38. Martins WS, Carmo WC, Longo HJ, Rosa TC, Rangel TF. 2013 SUNPLIN: Simulation with Uncertainty for Phylogenetic Investigations. *BMC Bioinformatics* 14, 324. (https://doi.org/10.1186/1471-2105-14-324)

39. Faith DP. 1992 Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. (https://doi.org/10.1016/0006-3207(92)91201-3)

40. Crisp MD, Laffan S, Linder HP, Monro A. 2001 Endemism in the Australian flora. *J. Biogeogr.* 28, 183–198. (https://doi.org/10.1046/j.1365-2699.2001.00524.x)

41. Shipley BR, McGuire JL. 2022 Interpreting and integrating multiple endemism metrics to identify hotspots for conservation priorities. *Biol. Conserv.* 265, 109403. (https://doi.org/10.1016/j.biocon.2021.109403)

42. CIESIN – Center for International Earth Science Information Network. 2018 *Documentation for the gridded population of the world, ver. 4 (GPWv4), revision 11 data sets*. NASA Socioeconomic Data and Applications Center (SEDAC) 4, 1–53. Accessed: 29 July 2025. (https://doi.org/10.7927/H49C6VHW)

43. Chen J, Gao M, Cheng S, Hou W, Song M, Liu X, Liu Y. 2022 Global 1 km × 1 km gridded revised real gross domestic product and electricity consumption during 1992–2019 based on calibrated nighttime light data. *Sci. Data* 9, 202. (https://doi.org/10.1038/s41597-022-01322-5)

44. Meijer JR, Huijbregts MAJ, Schotten KCGJ, Schipper AM. 2018 Global patterns of current and future road infrastructure. *Environ. Res. Lett.* 13, 064006. (https://doi.org/10.1088/1748-9326/aabd42)

45. Aranda SC, Gabriel R, Borges PAV, Azevedo EB, Lobo JM. 2011 Designing a survey protocol to overcome the Wallacean shortfall: a working guide using bryophyte distribution data on Terceira Island (Azores). *Bryologist* 114, 611. (https://doi.org/10.1639/0007-2745-114.3.611)

46. Chao A, Kubota Y, Zelený D, Chiu C, Li C, Kusumoto B, Yasuhara M, Thorn S, Wei C, Costello MJ, Colwell RK. 2020 Quantifying sample completeness and comparing diversities among assemblages. *Ecol. Res.* 35, 292–314. (https://doi.org/10.1111/1440-1703.12102)

47. Colwell RK, Chao A, Gotelli NJ, Lin SY, Mao CX, Chazdon RL, Longino JT. 2012 Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblage. *J. Plant Ecol.* 5, 3–21. (https://doi.org/10.1093/jpe/rtr044)

48. Hsieh TC, Ma KH, Chao A. 2020 iNEXT: interpolation and extrapolation for species diversity. R package version 2.0.20. (https://CRAN.R-project.org/package=iNEXT)

49. Cressie NAC. 1993 *Statistics for Spatial Data*. Wiley.

50. Diniz-Filho JAF, Rangel TFLVB, Bini LM. 2008 Model selection and information theory in geographical ecology. *Glob. Ecol. Biogeogr.* 17, 479–488. (https://doi.org/10.1111/j.1466-8238.2008.00395.x)

51. Kissling WD, Carl G. 2008 Spatial autocorrelation and the selection of simultaneous autoregressive models. *Glob. Ecol. Biogeogr.* 17, 59–71. (https://doi.org/10.1111/j.1466-8238.2007.00334.x)

52. Bivand RS, Pebesma E, Gómez-Rubio V. 2013 *Applied Spatial Data Analysis with R*. Springer Science & Business Media.

53. Almeida EA, Danforth BN. 2008 Phylogeny of colletid bees (Hymenoptera: Colletidae) inferred from four nuclear genes. *Mol. Phylogenet. Evol* 50, 290–309. (https://doi.org/10.1016/j.ympev.2008.09.028)

54. Almeida EAB, Pie MR, Brady SG, Danforth BN. 2011 Biogeography and diversification of colletid bees (Hymenoptera: Colletidae): emerging patterns from the southern end of the world. *J. Biogeogr.* 39, 526–544. (https://doi.org/10.1111/j.1365-2699.2011.02624.x)

55. Danforth BN, Conway L, Ji S. 2003 Phylogeny of Eusocial Lasioglossum Reveals Multiple Losses of Eusociality within a Primitively Eusocial Clade of Bees (Hymenoptera: Halictidae). *Syst. Biol.* 52, 23–36. (https://doi.org/10.1080/10635150390132687)

56. Gibbs J, Brady SG, Kanda K, Danforth BN. 2012 Phylogeny of halictine bees supports a shared origin of eusociality for Halictus and Lasioglossum (Apoidea: Anthophila: Halictidae). *Mol. Phylogenet. Evol.* 65, 926–939. (https://doi.org/10.1016/j.ympev.2012.08.013)

57. Pisanty G, Richter R, Martin T, Dettman J, Cardinal S. 2021 Molecular phylogeny, historical biogeography and revised classification of andrenine bees (Hymenoptera: Andrenidae). *Mol. Phylogenet. Evol.* 170, 107151. (https://doi.org/10.1016/j.ympev.2021.107151)

58. Wood TJ. 2025 Additions, corrections, and other changes to the hyper-diverse bee genus Andrena Fabricius, 1775 (Hymenoptera: Andrenidae). *Anim. Taxon. Ecol.* 71, 143–316. (https://doi.org/10.1556/1777.2025.00082)

59. Pereira FW, Gonçalves RB, Ramos KDS. 2021 Bee surveys in Brazil in the last six decades: a review and scientometrics. *Apidologie* 52, 1152–1168. (https://doi.org/10.1007/s13592-021-00894-2)

60. Chesshire PR, Fischer EE, Dowdy NJ, Griswold TL, Hughes AC, Orr MC, Ascher JS, Guzman LM, Hung KJ, Cobb NS, McCabe LM. 2023 Completeness analysis for over 3000 United States bee species identifies persistent data gap. *Ecography*, 2023, e06584. (https://doi.org/10.1111/ecog.06584)

61. Cornwell WK, Pearse WD, Dalrymple RL, Zanne AE. 2019 What we (don't) know about global plant diversity. *Ecography* 42, 1819–1831. (https://doi.org/10.1111/ecog.04481)

62. Guedes JJM, Diniz-Filho JAF, Moura MR. 2024 Macroecological correlates of Darwinian shortfalls across terrestrial vertebrates. *Biol. Lett.* 20, 20240216. (https://doi.org/10.1098/rsbl.2024.0216)

63. Tucker CM, Cadotte MW. 2013 Unifying measures of biodiversity: understanding when richness and phylogenetic diversity should be congruent. *Divers. Distrib*. 19, 845–854. (https://doi.org/10.1111/ddi.12087)

64. Meyer C, Kreft H, Guralnick R, Jetz W. 2015 Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* 6, 8221. (https://doi.org/10.1038/ncomms9221)

65. Wetterstrand KA. 2025 DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). (www.genome.gov/sequencingcostsdata). Accessed: 11 Oct. 2025.

66. Bratt S, Langalia M, Nanoti A. 2023 North-south scientific collaborations on research datasets: a longitudinal analysis of the division of labor on genomic datasets (1992–2021*). Front. Big Data* 6, 1054655. (https://doi.org/10.3389/fdata.2023.1054655)

67. Linck EB, Cadena CD. 2024 A latitudinal gradient of reference genomes. *Mol. Ecol.* e17551. (https://doi.org/10.1111/mec.17551)

68. Abreu ECT, Silva EL, Moura MR. 2025 Geopolitical impacts on the description of new terrestrial mollusc species. *Proc. R. Soc. B.* 292, 20251428. (https://doi.org/10.1098/rspb.2025.1428)

69. Nakamura G, Stabile BHM, Frateles LEF, Araujo ML, Neuhaus EB, Marinho MMF, Souza Leite M, Richter A, Ding L, Silva Freitas TM, Soares B, Graça WJ, Moura MR, Diniz-Filho JAF. 2025 The hidden biodiversity knowledge split in biological collections. *Proc. R. Soc. B. 292*, 20251045. (https://doi.org/10.1098/rspb.2025.1045)

70. Moura M, Carvalho R, Ceron K, Guedes J, Moroti M, Nakamura G. 2025 Local expertise anchors biodiversity documentation, but geopolitical power drives

parachute discovery. *Research Square* (preprint). (https://doi.org/10.21203/rs.3.rs-7724270/v1)

71. Oliveira U *et al*. 2016 The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Divers. Distrib.* 22, 1232-1244. (https://doi.org/10.1111/ddi.12489)

72. Hughes AC, Orr MC, Ma K, Costello MJ, Waller J, Provoost P, Yang Q, Zhu C, Qiao H. (2021) Sampling biases shape our view of the natural world. *Ecography*, 44, 1259-1269. (https://doi.org/10.1111/ecog.05926)

73. Blades B, Ronquillo C, Hortal J. 2025 Mobilisation of data from natural history collections can increase the quality and coverage of biodiversity information. *Ecol. Evol.* 15, e71139. (https://doi.org/10.1002/ece3.71139)

74. Castro-Souza RA, Tessarolo G, Stropp J, Diniz-Filho JAF, Ladle RJ, Szinwelski N, Hortal J, Sobral-Souza T. 2024 Mapping ignorance to uncover shortfalls in the knowledge on global Orthoptera distribution. *npj Biodivers.* 3, 22. (https://doi.org/10.1038/s44185-024-00059-1)